

# EPITOPE PROFILING VIA MIXTURE MODELING OF RANKED DATA

CRISTINA MOLLICA AND LUCA TARDELLA

**ABSTRACT.** We propose the use of probability models for ranked data as a useful alternative to a quantitative data analysis to investigate the outcome of bioassay experiments, when the preliminary choice of an appropriate normalization method for the raw numerical responses is difficult or subject to criticism. We review standard distance-based and multistage ranking models and in this last context we propose an original generalization of the *Plackett-Luce model* to account for the order of the ranking elicitation process. The usefulness of the novel model is illustrated with its maximum likelihood estimation for a real data set. Specifically, we address the heterogeneous nature of experimental units via model-based clustering and detail the necessary steps for a successful likelihood maximization through a hybrid version of the Expectation-Maximization algorithm. The performance of the mixture model using the new distribution as mixture components is compared with those relative to alternative mixture models for random rankings. A discussion on the interpretation of the identified clusters and a comparison with more standard quantitative approaches are finally provided.

DIPARTIMENTO DI SCIENZE STATISTICHE, SAPIENZA UNIVERSITÀ DI ROMA, PIAZZALE A. MORO 5, 00185 ROMA, ITALY

*E-mail address:* `cristina.mollica@uniroma1.it`

DIPARTIMENTO DI SCIENZE STATISTICHE, SAPIENZA UNIVERSITÀ DI ROMA, PIAZZALE A. MORO 5, 00185 ROMA, ITALY

*E-mail address:* `luca.tardella@uniroma1.it`

## 1. INTRODUCTION

Ranked data arise in several contexts, especially when objective and precise measurements of the phenomena of interest can be impossible or deemed unreliable and the observer gathers ordinal information in terms of orderings, preferences, judgments, relative or absolute ranking among competitors. Research fields where the analysis of ranked data are frequently required are the social and behavioral sciences, where studies often ask a sample of  $N$  people to rank a finite set of  $K$  items according to certain criteria, typically their personal preferences or attitudes. In marketing, political surveys or psychological experiments, items to rank can be consumer goods, political candidates or goals, words or topics considered to be more or less associated to a reference one according to the individual perception. Another typical context is sport, where teams, horses or cars compete, and the final outcome is a ranking among competitors. A detailed and well-structured reference monograph concerning ranking data analysis and modeling is (Marden, 1995).

---

*Key words and phrases.* Ranking data, Plackett-Luce model, Multistage ranking models, Mixture models, EM algorithm, Epitope mapping.

Version January 8, 2014.

Statistical analysis of observed rankings is less usual in experimental research, where the availability of (sometimes sophisticated) measuring devices allows to express phenomena of interest in terms of precise quantitative information. In this work we verified the usefulness of probability models for ranked data in an experimental study, where quantitative outcomes are indeed available but, for reasons due to numerical instability of the measurements and to the absence of universally accepted ways of rescaling the original data, one could instead investigate the ranking evidence. For this purpose we used a real data set from the Large Fragment Phage Display (LFPD) bioassay experiment described in (Gabrielli et al., 2013). Researchers set up a new promising technology in order to get further insights into the understanding of molecular recognition of the immune system via epitope mapping of the HER2 oncoprotein. They employed a sample of patients and recorded for each subject the binding level, expressed on quantitative scale, between antibodies and specific partially overlapping fragments of the HER2 oncoprotein. The sample was actually composed of three different groups according to the known breast cancer status. A preliminary exploratory analysis of the LFPD data showed differential outcome profiles for the three cancer-specific groups (Gabrielli et al., 2013). Hence, we assumed the sample as drawn from a heterogeneous, multimodal population and opted to describe it through a mixture modeling approach for the individual ranked binding sequences. Two well-studied probability distributions for ranked data, the distance-based and the Plackett-Luce model, and a new extension of the latter have been employed as mixture components and the resulting performances have been compared. Maximum likelihood estimates (MLE) have been obtained with the implementation of the Expectation-Maximization (EM) algorithm or with hybrid versions thereof.

This article is organized as follows: in Section 2 we define the notation and review the distance-based and the Plackett-Luce model. The presentation and motivation of our extended Plackett-Luce model follow in Section 2.3 and the MLE for a finite mixture is discussed in Section 3. The application to the LFPD data and the comparison of the novel model with alternative ranking probability distributions are detailed in Section 4, with an interpretation of the inferential findings. The article ends with conclusions and proposals for future developments in Section 5.

## 2. STATISTICAL MODELS FOR RANKED DATA

**2.1. Notation and basic definitions.** Before reviewing some of the approaches for the probabilistic modeling of ranked data, it is convenient to fix some notation. Formally, a *full* (or *complete*) *ranking* is a bijective mapping of a finite set  $I = \{i_1, \dots, i_K\}$  of labeled *items* into a set of *ranks*  $R = \{1, \dots, K\}$ , that is

$$\pi : I \rightarrow R.$$

With some abuse of notation, each item label will be identified with its subscript: instead of writing a ranking as  $\pi = (\pi(i_1), \dots, \pi(i_K))$ , we will simplify it as  $\pi = (\pi(1), \dots, \pi(K))$ . In this way positions refer to items and entries give the corresponding assigned ranks, which means that  $\pi(i)$  must be read as the rank attributed to the  $i$ -th item. The underlying convention is that if  $\pi(i) < \pi(i')$ , then item  $i$  is ranked higher than item  $i'$ , and hence preferred to it.

In the literature, one distinguishes a *full* from a *partial* (or *incomplete*) *ranking*, in which the rank assignment process is not completely carried out. This happens,

for instance, when a judge expresses only her first  $t$  preferences out of  $K$  items ( $t < K$ ), producing the so-called top- $t$  partial ranking. In the present context, the above restrictive definition for ranking is adopted, so that ties are not allowed due to injectivity and partial rankings are not contemplated because of surjectivity of the mapping  $\pi$ .

The inverse  $\pi^{-1} = (\pi^{-1}(1), \dots, \pi^{-1}(K))$  of a ranking  $\pi$  is called *ordering*. Positions of the components of  $\pi^{-1}$  refer to ranks and elements correspond to the items. Hence,  $\pi^{-1}(j)$  is the item ranked in the  $j$ -th position. In order to avoid confusion with  $\pi$ , we will henceforth make explicit use of the inverse function notation to denote the corresponding ordering  $\pi^{-1} : R \rightarrow I$ .

We denote with  $\mathcal{S}_K$  the set of all  $K!$  possible permutations. This special finite subset of  $\mathbb{R}^K$  is endowed with a composition operation such that two elements  $\pi$  and  $\sigma$  in  $\mathcal{S}_K$  may yield either a permutation of  $R$  or of  $I$ . In particular,  $\pi\sigma^{-1} = \pi \circ \sigma^{-1} = (\pi(\sigma^{-1}(1)), \dots, \pi(\sigma^{-1}(K)))$  indicates ranks under  $\pi$  of the items ranked  $1, \dots, K$  by  $\sigma$ , whereas  $\sigma^{-1}\pi = \sigma^{-1} \circ \pi = (\sigma^{-1}(\pi(1)), \dots, \sigma^{-1}(\pi(K)))$  gives items to which  $\sigma$  assigns ranks that  $\pi$  has attributed to items  $1, \dots, K$ .

When a judge proceeds from the elicitation of her best choice (rank 1) up to the worst one (rank  $K$ ), we have the so-called *forward ranking process*; the inverse ranking procedure is named *backward ranking process*. This formal definition has been originally introduced in (Fligner and Verducci, 1988) but, to our knowledge, the rank assignment scheme has not received an explicit consideration in a model setup in the attempt to improve the description of random ranked data. Obviously, any other order for the rank assignment process is admissible and potentially leads to different models. This aspect has inspired us to expand an existing and well-known parametric ranking model and employ such a new class in the analysis of the LFPD data, in order to verify whether and how the reference order can influence the inferential results and the final model-based clustering.

**2.2. Probability models for random rankings.** In this section we give a brief account of rank data modeling. For a more systematic review see (Marden, 1995).

The collection of all discrete distributions for random rankings can be identified with the whole  $(K! - 1)$ -dimensional simplex  $\mathcal{P}(\mathcal{S}_K)$ . This is equivalent to saying that a random ranking and its distribution can be denoted with  $\pi \sim P$ , where the set  $\{P \in \mathcal{P}(\mathcal{S}_K)\}$  can be regarded as the most general statistical model on rankings parameterized by  $K! - 1$  free parameters, i.e., the probabilities of each ordered sequence. This general form can be considered and named *saturated model* (SM). Within this very general class, a special role is played by the *uniform* or *null model* (UM), represented by the single flat distribution which assigns equal probability to each ranking, and by its opposites, the *degenerate models* (DM), which concentrate all the probability mass on a single ranking. Although the SM allows for the maximum degree of flexibility, it becomes intractable and cumbersome to interpret even with a relatively small number  $K$  of items, because of the fast-growing dimension of the ranking space. These practical limitations have motivated the introduction of simplifying assumptions on the ranking process, in order to deal with subsets of  $\mathcal{P}(\mathcal{S}_K)$ , and justify the wide assortment of restricted parametric models developed in the rank data theory.

2.2.1. *Distance-based models.* A fundamental class of parametric distributions is the so-called *distance-based model* (DB). Roughly speaking, the DB can be interpreted as the analogue of the normal distribution on the finite discrete space  $\mathcal{S}_K$  endowed with the group structure; in fact, it is an exponential location-scale model indexed by a discrete parameter  $\sigma \in \mathcal{S}_K$ , called *modal* or *central ranking*, and a non-negative real *concentration parameter*  $\lambda \in \mathbb{R}_+$ . Each distribution in a DB model has the following form

$$(2.1) \quad P(\pi|\sigma, \lambda) = \frac{1}{Z(\lambda)} e^{-\lambda d(\pi, \sigma)} \quad \pi \in \mathcal{S}_K,$$

where  $Z(\lambda) = \sum_{\pi \in \mathcal{S}_K} e^{-\lambda d(\pi, \sigma)}$  is the normalization constant and  $d$  is a metric on  $\mathcal{S}_K$ . The probability mass function in (2.1) attains its maximum at  $\pi = \sigma$  and decreases as the distance from  $\sigma$  increases. Under (2.1) rankings at the same distance from the modal sequence  $\sigma$  are equally probable. The central ranking  $\sigma$  expresses the so-called *global consensus* in the population, whereas the concentration/precision parameter  $\lambda$  calibrates the effect of  $d$  on the probability of the ranking; the higher the value of  $\lambda$ , the more concentrated the distribution around its mode. Hence, when  $\lambda \rightarrow +\infty$ , equation (2.1) becomes the DM at  $\pi = \sigma$ ; conversely, when  $\lambda = 0$  it turns out to be the UM.

Changing the distance measure  $d$  in (2.1), one can define different families of parametric distributions for ranked data. Formally, a function  $d : \mathcal{S}_K \times \mathcal{S}_K \rightarrow \mathbb{R}_+$  is a distance between rankings if it satisfies the usual three properties of a metric (identity, symmetry and triangle inequality) and the additional fourth condition of *right-invariance*, that is for all  $\pi, \xi, \delta \in \mathcal{S}_K$

$$(2.2) \quad d(\pi, \xi) = d(\pi\delta^{-1}, \xi\delta^{-1}).$$

Condition (2.2) guarantees the desirable property of invariance of  $d$  w.r.t. arbitrary relabeling of items. Examples of metrics for rankings are:

the *Kendall distance*

$$d_K(\pi, \xi) = \sum_{1 \leq i < i' \leq K} I_{[(\pi(i) - \pi(i'))(\xi(i) - \xi(i')) < 0]},$$

which counts the number of pairwise disagreements, i.e., the pairs of items with relative discordant order under  $\pi$  and  $\xi$ . It is also equal to the minimum number of adjacent transpositions needed to transform  $\pi^{-1}$  into  $\xi^{-1}$ ;

the *Spearman distance*  $d_S(\pi, \xi) = \sum_{i=1}^K [\pi(i) - \xi(i)]^2$ ;

the *Spearman Footrule*  $d_F(\pi, \xi) = \sum_{i=1}^K |\pi(i) - \xi(i)|$ ;

the *Cayley distance*  $d_C(\pi, \xi) = K - C(\pi^{-1}\xi)$ , where  $C(\eta)$  is the number of cycles in  $\eta$ , corresponding to the minimum number of arbitrary transpositions required to convert  $\pi^{-1}$  into  $\xi^{-1}$ .

The reader is referred to (Critchlow, 1985) for a more complete and detailed description of the metrics on rankings.

The computation of  $Z(\lambda)$  can be computationally demanding as it requires the summation over all possible rankings. As advised by (Fligner and Verducci, 1986), one way to derive a simpler expression for  $Z(\lambda)$  is to consider its relation with the moment generating function of the random variable  $D(\pi, \sigma)$  under the UM on  $\mathcal{S}_K$ . In the wide variety of distances, only some specific ones lead to a closed form expression for  $Z(\lambda)$ . Hence, in performing a statistical analysis of ranked data one should balance between interpretation purposes, choosing the  $d$  which

best accommodates the problem at hand, and computational feasibility. For our application to the LFPD data we employed the Kendall distance  $d_K$ .

**2.2.2. The Plackett-Luce models and related extensions.** The *Plackett-Luce model* (PL) is a very popular parametric family for random ranking. Its name arises from both contributions supplied by (Luce, 1959), whose monograph provides an in-depth theoretical description of the individual choice behavior with a general axiom, and (Plackett, 1975), who derived this model in the context of horse races. Its probabilistic expression moves from the decomposition of the ranking process in independent stages, one for each rank that has to be assigned, combined with the underlying assumption of standard forward procedure on the ranking elicitation. In fact, a ranking can be elicited through a series of sequential comparisons in which a single item is preferred to all the remaining ones and after being selected is removed from the next comparisons. For this reason, the PL is said to belong to the family of *multistage ranking models*. Specifically, the PL probability distribution is completely specified by the so-called *support parameter* vector  $\underline{p} = (p_1, \dots, p_K)$ , where  $p_i > 0$  for all  $i = 1, \dots, K$  and  $\sum_{i=1}^K p_i = 1$ . Note that in the present formulation the parameters are constrained to add up to one to avoid non-identifiability due to possible multiplication with an arbitrary positive constant. The generic parameter  $p_i$  expresses the probability that item  $i$  is selected at the first stage of the ranking process and hence preferred among all other items. The probability to choose item  $i$  at lower preference levels  $t > 1$  is proportional to its support value  $p_i$ . Taking into account that the set of available items in the sequence of random selections is reduced by one element after each step, the computation of the choice probabilities at each stage for the assignment of the actual rank requires suitable normalization of the support probabilities w.r.t. the set of remaining items at that stage. It follows that under the PL the probability of the random ordering  $\pi^{-1}$  is

$$(2.3) \quad \mathbf{P}(\pi^{-1}|\underline{p}) = \prod_{t=1}^K \frac{p_{\pi^{-1}(t)}}{\sum_{v=t}^K p_{\pi^{-1}(v)}} \quad \pi^{-1} \in \mathcal{S}_K.$$

The vase model metaphor originally introduced by (Silverberg, 1980) is an alternative way to interpret the random stage-wise item selections and a useful representation of the PL to understand its extensions developed in the literature (see (Marden, 1995) for a review). Let us consider a vase containing item-labelled balls such that the vector  $\underline{p}$  expresses the starting composition of the vase. The vase differs from an urn simply because the former contains an infinite number of balls in order to allow continuous values of the proportions. At the first stage one draws a ball and ranks the corresponding item first. At the second stage one draws another ball from the vase: if its label is different from  $\pi^{-1}(1)$  one assigns rank 2 to the corresponding item, otherwise the ball is put back and one makes drawings until a distinct item is chosen and then ranked second. The multistage experiment ends when there is only one item not yet selected and this is automatically ranked last. The probability of a generic sequence of drawings turns out to be (2.3). In such a scheme the vase configuration is constant over all stages and interactions among items are not contemplated. A first attempt to generalize this basic scheme consists in retaining the absence of item interactions but letting the vase composition vary among stages, as formalized in (Silverberg, 1980). In this model setting the support parameters become stage-dependent, that is  $p_{ti}$  for  $t = 1, \dots, (K-1)$  and

$i = 1, \dots, K$ . Setting the special form  $p_{ti} = p_i^{\alpha_t}$  one obtains the *Benter model* (BM) introduced by (Benter, 1994), where the parameter vector  $\underline{\alpha} = (\alpha_1, \dots, \alpha_K)$  with  $0 \leq \alpha_t \leq 1$  for all  $t = 1, \dots, K$  is named *dampening parameter* and accommodates for the possible different degree of accuracy the choice at each selection stage is made with. The PL corresponds to the BM with  $\alpha_t = \alpha = 1$  for all  $t = 1, \dots, K$ . Relaxing also the non-interaction hypothesis, meaning that the vase composition at each stage relies on the previous selected items, (Plackett, 1975) defined a hierarchy of further extensions of the PL. They are referred to in (Marden, 1995) as *Lag L models*, where  $L = 0, \dots, K - 2$  indicates that the vase at stage  $t$  depends on the previous choices only through the last  $L$  selected items  $\{\pi^{-1}(t-L), \dots, \pi^{-1}(t-1)\}$ . The Lag 0 model coincides with the ordinary PL. The Lag 1 model is such that at each choice step  $t$  the vase depends only on the item  $\pi^{-1}(t-1)$ . In general, the total number of parameters in the Lag  $L$  model is given by  $K(K-1) \cdots (K-L) - 1$ , thus the  $L = K - 2$  model corresponds to the SM.

**2.3. Novel extension of the Plackett-Luce model.** In this section we introduce an original proposal to generalize the standard PL. Multistage ranking models previously reviewed implicitly suppose that preferences are expressed with the canonical forward procedure, proceeding with the assignment of the first rank up to the last one. This is just a preliminary assumption and other reference orders can be contemplated but, to our knowledge, this aspect has not been addressed in the literature. Indeed, even the individual experience in choice problems suggests the plausibility of alternative paths for the ranking elicitation. For example, one can think of situations where the judge has a clearer perception about her most- and least-liked items first but only a vaguer idea relative to middle ranks; alternatively again the ranker can build up her best alternatives following an exclusion process starting with the final position, which would be described by a backward model. Besides the motivation to characterize typical behaviors in real choice/selection problems, we can also aim at obtaining a more flexible tool in order to improve the description of observed phenomena collected in the form of ordered data. All these intuitive and practical instances make the forward hypothesis too restrictive when approaching a flexible inferential analysis of a ranking data set. Hence, we propose to extend the PL in this way: rather than fixing *a priori* the stepwise order leading the judge to her final ranked sequence, we would like to represent it with a specific free parameter  $\rho \in \mathcal{S}_K$  in the model and let data guide inference about the reference order followed in the rank assignment scheme. Such an approach would also alleviate the asymmetry toward ranks assigned at the extreme (the first and the last) stages of the ranking procedure, which by nature affects the PL with hypothesized known reference order. It turns out that the reference order  $\rho = (\rho(1), \dots, \rho(K))$  is the result of a bijection between the stage set  $S$  and the rank set  $R$ , i.e.,

$$\rho : S \rightarrow R,$$

where the entry  $\rho(t)$  indicates the rank attributed at the  $t$ -th stage of the ranking process. Then,  $\rho$  identifies a discrete parameter taking values in  $\mathcal{S}_K$ . The composition of an ordering  $\pi^{-1}$  with a reference order  $\rho$  yields

$$\eta^{-1} = \pi^{-1}\rho,$$

the sequence listing the items selected at each stage. This means that  $\eta^{-1}(t) = \pi^{-1}(\rho(t))$  is the item chosen at step  $t$  and receiving rank  $\rho(t)$ . The probability of a

random ordering under the *extended Plackett-Luce model* can be written as

$$(2.4) \quad \mathbf{P}_{EPL}(\pi^{-1}|\rho, \underline{p}) = \mathbf{P}_{PL}(\pi^{-1}\rho|\underline{p}) = \prod_{t=1}^K \frac{p_{\pi^{-1}(\rho(t))}}{\sum_{v=t}^K p_{\pi^{-1}(\rho(v))}} \quad \pi^{-1} \in \mathcal{S}_K,$$

where the additional discrete parameter  $\rho$  acts directly on the right of the generated outcome of a standard PL. Hereafter we will shortly refer to (2.4) as  $EPL(\rho, \underline{p})$ . The vector  $\underline{p}$  continues to denote the support parameters with the probabilities for each item to be selected at the first stage and receiving rank given by the first entry in  $\rho$ . Obviously, the standard PL is a special case of the EPL, obtained setting  $\rho$  equal to the identity permutation  $e = (1, 2, \dots, K)$ . Similarly, when  $\rho = (K+1) - e$  one has the backward PL.

From a theoretical point of view, (2.4) is a proper generalization of the (2.3) if and only if such a new class covers a wider portion of the SM, i.e., if the novel EPL allows to describe additional probability functions that can not be derived with any parameter specification from the PL. In other words, one should give a formal proof concerning the existence of a ranking distribution, generated by the new EPL, which does not belong to the standard PL family. Such a proof is given in the Appendix. In section 3.2 we describe in detail the MLE of such a new model.

**2.4. Finite mixture modeling for ranked data.** One of the nice formal properties satisfied by the DB (2.1) is strong unimodality, meaning that the probability decreases as the distance from the modal ranking increases, see (Marden, 1995). On the other hand strong unimodality is expected to be violated in real data, especially when the sample composition is heterogenous w.r.t. factors related to the ranking elicitation. A well-established statistical tool to address inference in the presence of unobserved heterogeneity is given by the finite mixture approach. A *finite mixture model* assumes that the population consists of a finite number  $G$  of subpopulations. In this setting the probability of observing the ranking  $\pi_s$  for the  $s$ -th unit is

$$f(\pi_s) = \sum_{g=1}^G \omega_g f_g(\pi_s) \quad \pi_s \in \mathcal{S}_K,$$

where  $f_g(\cdot)$  denotes the  $g$ -th *component* of the mixture, i.e., the statistical distribution of data within the  $g$ -th group and  $\omega_g$  is the probability for the  $s$ -th observation to belong to the  $g$ -th group. The membership probabilities  $\underline{\omega} = (\omega_1, \dots, \omega_G)$  are usually termed *weights* of the mixture components. Mixture components are often modeled with members of the same parametric family, that is,  $f_g(\cdot) = f(\cdot|\theta_g) \in \{f(\cdot|\theta) : \theta \in \Theta\}$  for all  $g = 1, \dots, G$  and thus they are identified by the group-specific parameter  $\theta_g$ . For a more extensive introduction to finite mixture models the reader can refer to (McLachlan and Peel, 2000). In the ranking literature one can find several recent mixture model applications to make the ranked data modeling more flexible and account for unobserved heterogeneity. For example, (Murphy and Martin, 2003) analyzed the popular 1980 APA (American Psychological Association) presidential election data set, in particular the sub-data set of complete rankings, with a mixture of distance-based models. They aimed at inquiring voters' orientation towards candidates within the electorate, assessing the possible adequacy to incorporate a noise component (UM) in the mixture. Such a component, in fact, could collect outliers and/or observations characterized by untypical preference profiles with a possible final improvement of model fitting. A similar method

was adopted in other preference studies. (Gormley and Murphy, 2006) fitted a mixture of PL to the 2000 CAO (Central Applicant Office) data set to investigate motivations driving Irish college applicants in the degree course choice. (Gormley and Murphy, 2008a) applied a mixtures of both PL and BM to infer the structure of the Irish political electorate and characterize voting blocks. In subsequent works the same authors attempted to extend such an approach in different directions (for further details see (Gormley and Murphy, 2008b) and (Gormley and Murphy, 2009)).

In section 4.2 we present our application of mixture models for ranking to data originated from the LFPD bioassay experiment. For the analysis of the LFPD data set we implemented different mixture models, adopting as mixture components elements from the following parametric families:

- DB with  $d = d_K$ ;
- PL with known forward and backward reference order;
- our novel EPL.

DB and PL represent two of the most frequently used distributions for inferring ranking data and both parameterizations allow a clear interpretation: in the former, the central ranking summarizes the profile of the population in assessing the orderings of the items and the concentration parameter expresses how representative the modal ranking is; in the latter, the higher the item support parameter value, the greater the probability for that item to be preferred at each selection level. For the argument on the choice of the EPL in the analysis of the LFPD data, the reader is referred to section 4.2.

### 3. INFERRING RANKING MODELS

**3.1. MLE of the mixture of distance-based models.** We briefly summarize here the fundamental steps to derive the MLE for a mixture of DB with  $d = d_K$  when a sample  $\underline{\pi} = (\pi_1, \dots, \pi_s, \dots, \pi_N)$  is available. We basically reproduce the algorithm described in (Murphy and Martin, 2003). Let  $\underline{z}_s = (z_{s1}, \dots, z_{sG})$  be the latent variable indicating the individual component membership such that

$$z_{sg} = \begin{cases} 1 & \text{if the } s\text{-th unit belongs to the } g\text{-th group,} \\ 0 & \text{otherwise,} \end{cases}$$

for  $s = 1, \dots, N$ . From (2.1) it follows that the complete log-likelihood can be written as

$$l_C(\underline{\sigma}, \underline{\lambda}, \underline{\omega}, \underline{z}) = \sum_{s=1}^N \sum_{g=1}^G z_{sg} [\log \omega_g - \lambda_g d_K(\pi_s, \sigma_g) - \log Z(\lambda_g)],$$

where  $\underline{\omega}$  and  $\underline{\lambda}$  are vectors representing, respectively, the prior group membership probabilities and the component-specific concentration parameters, whereas  $\underline{\sigma}$  is a  $G \times K$  matrix, whose rows indicate the central rankings of the mixture components. To derive parameter estimates the EM algorithm can be implemented; it represents the major scheme to address the inferential analysis in the presence of missing data (Dempster et al., 1977). For the present model the EM algorithm consists of the following steps:

Initialization: set initial values  $\underline{\sigma}^{(0)}, \underline{\lambda}^{(0)}, \underline{\omega}^{(0)}$  for the parameters to be estimated (we used random starting values).



E-step: at iteration  $l + 1$  compute

$$\hat{z}_{sg}^{(l+1)} = \frac{\omega_g^{(l)} \mathbf{P}_{DB}(\pi_s | \sigma_g^{(l)}, \lambda_g^{(l)})}{\sum_{g'=1}^G \omega_{g'}^{(l)} \mathbf{P}_{DB}(\pi_s | \sigma_{g'}^{(l)}, \lambda_{g'}^{(l)})},$$

for  $s = 1, \dots, N$  and  $g = 1, \dots, G$ , which is the current estimate of the posterior probability that subject  $s$  belongs to the  $g$ -th component;

M-step: at iteration  $l + 1$  compute

$$\omega_g^{(l+1)} = \sum_{s=1}^N \frac{\hat{z}_{sg}^{(l+1)}}{N},$$

$$\sigma_g^{(l+1)} = \arg \min_{\sigma} \sum_{s=1}^N \hat{z}_{sg}^{(l+1)} d(\pi_s, \sigma),$$

and determine  $\lambda_g^{(l+1)}$  as the solution of

$$\frac{K e^{-\lambda}}{1 - e^{-\lambda}} - \sum_{j=1}^K \frac{j e^{-j\lambda}}{1 - e^{-j\lambda}} = \frac{\sum_{s=1}^N \hat{z}_{sg}^{(l+1)} d(\pi_s, \sigma_g^{(l+1)})}{\sum_{s=1}^N \hat{z}_{sg}^{(l+1)}},$$

for  $g = 1, \dots, G$ . We run the algorithm with a suitably large number of different starting values to address the issue of local maxima.

**3.2. MLE of the mixture of Extended Plackett-Luce models.** As mentioned before, the conventional forward PL is a reduction of the wider family of EPL distributions obtained setting the reference order parameter  $\rho$  equal to the identity permutation  $e$ . It follows that the estimation procedure for the mixture of PL can be easily deduced from the one derived for the mixture of EPL with all known reference orders  $\rho_g = e$ . However, explicit estimation formulas for this special case can be found in (Gormley and Murphy, 2006). In this section we restrict ourselves to give inferential details only for the extended model, starting with the simpler case of homogenous population ( $G=1$ ).

Postulating the  $\text{EPL}(\rho, \underline{p})$  as the underlying mechanism generating the observed orderings  $\underline{\pi}^{-1} = (\pi_1^{-1}, \dots, \pi_N^{-1})$ , the log-likelihood has the following expression

$$\begin{aligned} l(\rho, \underline{p}) &= \sum_{s=1}^N \sum_{t=1}^K \left[ \log \frac{p_{\pi_s^{-1}(\rho(t))}}{\sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}} \right] \\ (3.1) \quad &= \sum_{s=1}^N \sum_{t=1}^K \left[ \log p_{\pi_s^{-1}(\rho(t))} - \log \left( \sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))} \right) \right] \\ &= N \sum_{i=1}^K \log p_i - \sum_{s=1}^N \sum_{t=1}^K \log \left( \sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))} \right). \end{aligned}$$

Note that in order to find MLE solutions, the direct maximization of the log-likelihood w.r.t. the  $p$ 's is made arduous by the presence of the annoying term  $\log \left( \sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))} \right)$ . So, we derived the estimation formula for the support parameters borrowing the approach detailed in (Hunter, 2004) and based on the Minorization/Maximization (MM) algorithm. Such an iterative optimization method is reviewed in general in (Lange et al., 2000) and (Hunter and Lange, 2004), whereas (Hunter, 2004) discusses the specific application of the MM algorithm

for the estimation of the PL. The basic idea consists of performing the optimization step for the  $p$ 's on a surrogate objective function rather than on (3.1). The surrogate is obtained by exploiting the strict convexity of  $-\log\left(\sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}\right)$  and in particular the supporting hyperplane property for convex functions. From Taylor's linear expansion, in fact, one has

$$-\log\left(\sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}\right) \geq 1 - \log\left(\sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}^{(l)}\right) - \sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))} / \sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}^{(l)},$$

and disregarding terms not depending on  $\underline{p}$  the minorizing auxiliary objective function can be written as

$$(3.2) \quad q = N \sum_{i=1}^K \log p_i - \sum_{s=1}^N \sum_{t=1}^K \frac{\sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}}{\sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}^{(l)}}.$$

As emphasized by (Hunter, 2004), the convenience of optimizing the more tractable (3.2) in place of (3.1) is the possibility to estimate each support parameter  $p_i$  separately. Furthermore, the iterative maximization of  $q$  returns a sequence  $\underline{p}^{(1)}, \underline{p}^{(2)}, \dots$  that is guaranteed to converge at least to a local maximum of the original objective function. Thus, we can differentiate w.r.t. each  $p_i$  and get

$$(3.3) \quad \frac{\partial q}{\partial p_i} = \frac{N}{p_i} - \sum_{s=1}^N \sum_{t=1}^K \frac{\delta_{sti}^{(l)}}{\sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}^{(l)}},$$

where

$$\delta_{sti}^{(l)} = \begin{cases} 1 & \text{if } i \in \{\pi_s^{-1}(\rho^{(l)}(t)), \dots, \pi_s^{-1}(\rho^{(l)}(K))\}, \\ 0 & \text{otherwise,} \end{cases}$$

corresponds to the binary indicator for the event that item  $i$  is still available at stage  $t$  or, equivalently, that is not selected by unit  $s$  before stage  $t$ . Notice that the binary array has a superscript because of the dependence on the  $\rho = \rho^{(l)}$  available at the current iteration. Equating (3.3) to zero, the updating rule at the current iteration for  $p_i$  is

$$p_i^{(l+1)} = \frac{N}{\sum_{s=1}^N \sum_{t=1}^K \frac{\delta_{sti}^{(l)}}{\sum_{v=t}^K p_{\pi_s^{-1}(\rho^{(l)}(v))}^{(l)}}} \quad i = 1, \dots, K.$$

The update of the reference order parameter is obtained using the original log-likelihood as follows

$$(3.4) \quad \rho^{(l+1)} = \arg \min_{\rho} \sum_{s=1}^N \sum_{t=1}^K \log \left( \sum_{v=t}^K p_{\pi_s^{-1}(\rho(v))}^{(l+1)} \right).$$

Solving (3.4) with a global search in  $\mathcal{S}_K$  is prohibitive when  $K$  is large, as in our application to the LFPD data. So, we implemented a local search similarly to the method suggested by (Busse et al., 2007) and constrained the optimization within a fixed Kendall distance from the current estimate for the reference order  $\rho^{(l)}$ .

Now we relax the hypothesis of homogeneous population and consider a more flexible mixture model with EPL components, discussing the related inferential issues. If we assume our random sample  $\underline{\pi}^{-1}$  drawn from a mixture of EPL, the probability of the generic ordering is written as the average of its probability under

each sub-population weighted with the corresponding mixture component weight, i.e.,

$$\begin{aligned}\mathbf{P}(\pi_s^{-1} | \underline{\rho}, \underline{p}, \underline{\omega}) &= \sum_{g=1}^G \omega_g \mathbf{P}_{EPL}(\pi_s^{-1} | \rho_g, \underline{p}_g) \\ &= \sum_{g=1}^G \omega_g \prod_{t=1}^K \frac{p_{g\pi_s^{-1}(\rho_g(t))}}{\sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g(v))}}.\end{aligned}$$

Hence, the observed log-likelihood turns out to be

$$l(\underline{\rho}, \underline{p}, \underline{\omega}) = \sum_{s=1}^N \log \left[ \sum_{g=1}^G \omega_g \prod_{t=1}^K \frac{p_{g\pi_s^{-1}(\rho_g(t))}}{\sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g(v))}} \right].$$

Augmenting data with the missing individual group membership indicator  $\underline{z}_s = (z_{s1}, \dots, z_{sG})$ , one obtains the following expression for the complete log-likelihood

$$\begin{aligned}l_C(\underline{\rho}, \underline{p}, \underline{\omega}, \underline{z}) &= \log \mathbf{P}(\pi^{-1}, \underline{z} | \underline{\rho}, \underline{p}, \underline{\omega}) = \log \prod_{s=1}^N \mathbf{P}(\pi_s^{-1} | \underline{z}_s, \underline{\rho}, \underline{p}) \mathbf{P}(\underline{z}_s | \underline{\omega}) \\ &= \log \prod_{s=1}^N \prod_{g=1}^G \left[ \omega_g \prod_{t=1}^K \frac{p_{g\pi_s^{-1}(\rho_g(t))}}{\sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g(v))}} \right]^{z_{sg}} \\ &= \sum_{s=1}^N \sum_{g=1}^G z_{sg} \left[ \log \omega_g + \sum_{i=1}^K \log p_{gi} - \sum_{t=1}^K \log \left( \sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g(v))} \right) \right].\end{aligned}$$

In the EM algorithm the maximization problem is transferred on the the expectation of the  $l_C$  w.r.t. the posterior distribution of the latent variables  $\underline{z}$  represented by  $\hat{\underline{z}}$ , that is

$$Q = \mathbf{E}[l_C | \pi^{-1}, \underline{\rho}, \underline{p}, \underline{\omega}] = \sum_{s=1}^N \sum_{g=1}^G \hat{z}_{sg} \left[ \log \omega_g + \sum_{i=1}^K \log p_{gi} - \sum_{t=1}^K \log \left( \sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g(v))} \right) \right],$$

where

$$\hat{z}_{sg}^{(l+1)} = \frac{\omega_g^{(l)} \mathbf{P}_{EPL}(\pi_s^{-1} | \rho_g^{(l)}, \underline{p}_g^{(l)})}{\sum_{g'=1}^G \omega_{g'}^{(l)} \mathbf{P}_{EPL}(\pi_s^{-1} | \rho_{g'}^{(l)}, \underline{p}_{g'}^{(l)})},$$

for  $s = 1, \dots, N$  and  $g = 1, \dots, G$ . Similarly to (Gormley and Murphy, 2006) we combined the EM with the MM algorithm into a hybrid version of the former called EMM algorithm using the following minorizing surrogate function

$$Q \geq q = \sum_{s=1}^N \sum_{g=1}^G \hat{z}_{sg} \sum_{i=1}^K \log p_{gi} - \sum_{s=1}^N \sum_{g=1}^G \hat{z}_{sg} \sum_{t=1}^K \frac{\sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g(v))}}{\sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g(v))}^{(l)}}.$$

Thus, differentiating

$$(3.5) \quad \frac{\partial q}{\partial p_{gi}} = \frac{\sum_{s=1}^N \hat{z}_{sg}}{p_{gi}} - \sum_{s=1}^N \hat{z}_{sg} \sum_{t=1}^K \frac{\delta_{stig}}{\sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g(v))}^{(l)}}$$

and equating (3.5) to zero, the updating rule for  $p_{gi}$  at the current iteration is

$$p_{gi}^{(l+1)} = \frac{\sum_{s=1}^N \hat{z}_{sg}^{(l+1)}}{\sum_{s=1}^N \hat{z}_{sg}^{(l+1)} \sum_{t=1}^K \frac{\delta_{stig}^{(l)}}{\sum_{v=t}^K p_{g\pi_s^{-1}(\rho_g^{(l)}(v))}^{(l)}}}$$

for  $g = 1, \dots, G$  and  $i = 1, \dots, K$  with

$$\delta_{stig}^{(l)} = \begin{cases} 1 & \text{if } i \in \{\pi_s^{-1}(\rho_g^{(l)}(t)), \dots, \pi_s^{-1}(\rho_g^{(l)}(K))\}, \\ 0 & \text{otherwise,} \end{cases}$$

indicating if, under the group-specific reference order  $\rho_g$ , the unit  $s$  has not extracted the  $i$ -th item before stage  $t$ , and hence if at that step it still belongs to the set of available alternatives or not. The update for the reference orders for each subgroup is

$$\rho_g^{(l+1)} = \arg \min_{\rho} \sum_{s=1}^N \hat{z}_{sg}^{(l+1)} \sum_{t=1}^K \log \left( \sum_{v=t}^K p_{g\pi_s^{-1}(\rho(v))}^{(l+1)} \right) \quad g = 1, \dots, G.$$

As in the case  $G = 1$  in (3.4), the above minimization is performed locally. The M-step ends with the update of the mixture weights, computed as the posterior proportions of sample units belonging to each group

$$\omega_g^{(l+1)} = \frac{\sum_{s=1}^N \hat{z}_{sg}^{(l+1)}}{N} \quad g = 1, \dots, G.$$

**3.3. Algorithm convergence and model selection.** We conducted MLE inference for DB and EPL mixture models relying on the EM algorithm and on a hybrid version thereof. We developed a suite of functions written in R language (R Core Team, 2012) which are available upon request from the first author. In these estimation procedures the log-likelihood is iteratively maximized until convergence is achieved. As suggested by (McLachlan and Peel, 2000), the Aitken acceleration criterion has been employed to assess convergence, in place of the standard lack of progress criterion based on the absolute/relative increment of the log-likelihood. For a discussion on the relative merits of the Aitken acceleration criterion and other related proposals, see (McNicholas et al., 2010).

Another crucial issue in a mixture modeling setting is the choice of the number of components. In the statistical literature this problem is addressed with several criteria; we opted for the popular *Bayesian Information Criterion*

$$\text{BIC} = -2l(\hat{\theta}_{ML}) + \nu \log N,$$

where  $l(\hat{\theta}_{ML})$  is the maximized log-likelihood and  $\nu$  is the number of free parameters. The BIC, introduced by (Schwarz, 1978), is a measure which balances between two conflicting goals typically aimed at when fitting a statistical model: good fit and parameter parsimony, where the latter is modulated through the penalty term. In the presence of competing mixture models, the one associated with the lowest value of the BIC is preferred. In the next section we detail MLE results derived from alternative mixture models fitted to the LFPD data set.

#### 4. STATISTICAL ANALYSIS OF LFPD DATA

**4.1. The LFPD data set.** Our investigation is motivated by a real data set coming from a new technology for epitope mapping of the binding between the antibodies present in a biological tissue and a target protein. The biological foundation of the experiment is detailed in (Gabrielli et al., 2013) and consists of repeated binding measurements of human blood exposed to  $K = 11$  partially overlapping fragments of the HER2 oncoprotein, denoted sequentially by Hum 1,  $\dots$ , Hum 11 (see Figure 1). Researchers were originally interested in testing the validity of their innovative biotechnology which consists in a new way of isolating protein fragments without losing the conformational structure of the protein portions. To achieve this goal they employed a phage as a vector for hosting each protein fragment. Then they compared the binding outcome detected on each of the 11 fragments via a standard Enzyme-Linked Immunosorbent Assay (ELISA) with the whole protein (Hum 12) and the empty vector (Hum 13), used respectively as positive and negative controls (see Figure 1).

[Figure 1 about here.]

They first checked with monoclonal antibodies that the expected binding at some specific fragment was actually detected. Then they gathered  $N = 67$  samples of human blood taken from three different disease groups: i) HD = healthy patients, ii) EBC = patients diagnosed with breast cancer at an early stage, iii) MBC = patients diagnosed with metastatic breast cancer. Binding outcomes from the ELISA experiment have been detected by a laser scanner so that the binding intensities have been measured and recorded in terms of absorbance levels in nanometers (nm). In the next section we motivate our statistical analysis of the LFPD data based on the ordinal information, rather than on the original quantitative scale measurements.

**4.2. Mixture models for the LFPD data.** The original raw absorbance data derived from the LFPD experiment were somehow wildly fluctuating and looked indeed very heterogeneous as apparent in Figure 2.

[Figure 2 about here.]

However there were certainly some manifest peaks corresponding to recurrent fragments, especially high for some patients, most frequently those diagnosed with cancer. It is also apparent that the individual absorbance profiles are measured at different mean levels for different patients and with different variability. A simple logarithmic transformation and recentering w.r.t. the individual mean were performed providing some more stable evidence of differential profiles among groups. However there are some specific profiles which seem pretty much overlapped, among different subgroups although with some different overall pattern (Figure 3).

[Figure 3 about here.]

Since data emerged from the development of an innovative technology, miscalibrations or inaccuracies of the measuring device may occur and/or subject-specific characteristics may alter somehow the observed numerical outcome, making it more difficult to adjust the statistical analysis based on raw or ad-hoc pre-processed data. Unfortunately, for this kind of data a consolidated and fully-shared normalization technique is lacking. For all these reasons, rather than basing our analysis on the quantitative output of the LFPD study, we verified the possible usefulness of the ranking profiles as a more robust and unambiguously-defined evidence, capable to

capture and characterize the sample heterogeneity w.r.t. the disease status and specific characteristic profiles of each subgroup. Hence, we first derived ordered sequences ranking the absorbance levels of the individual protein fragments taken in decreasing order (Rank 1=highest value, Rank  $K$ =lowest value). We performed a simple exploratory analysis by cancer status computing both the  $K \times K$  first-order marginal matrices  $\hat{M}$ , where the generic entry  $\hat{M}_{ij}$  indicates the observed relative frequency that item  $i$  is ranked  $j$ -th, and the so-called Borda orderings  $\bar{\pi}^{-1}$ , listing items taken in order from the highest to the lowest mean rank. These matrices are displayed as image plots in Figure 4,

[Figure 4 about here.]

together with the Borda sequences on the bottom of each panel. The color intensity is an increasing function of the corresponding observed frequency. The analysis of the first-order marginals matrices suggests that very often some protein fragments are associated with lower ranks, as pointed out by the presence of darker rectangles in correspondence of bottom positions. This constantly occurs for all disease subgroups with Hum 10 but some interesting differential evidence emerges from EBC subjects with Hum 5 and 6, from MBC with Hum 2 and 13 and also from HD patients with Hum 9 (Figure 4). Such a precious discriminant information could be better captured by our EPL. To validate this claim we fitted both the PL and the new EPL to the three disease subgroups separately. For the former we used known orders, alternatively forward (PL- $\rho_1$ ) and backward (PL- $\rho_2$ ), whereas for the latter the reference order is a parameter to be estimated. Estimation performances are shown in terms of BIC values in Table 1 and reveal that the fit of the EPL is better or at most comparable with those relative to the PL with fixed reference orders. The interest in relaxing the canonical forward assumption is supported also by the BIC values for the PL- $\rho_2$ , showing such a model constantly outperforms the PL- $\rho_1$  when fitted to HD and MBC subjects. These BIC results represent a strong evidence motivating the need of an extension of the PL.

[Table 1 about here.]

Subsequently we considered a more comprehensive analysis in a mixture model setting. With this approach we aimed at:

- addressing the heterogenous nature of the LFPD data using the evidence provided by the orderings of absorbance levels;
- assessing if and how the path in the sequential ranking process can have an impact on the final model-based classification of experimental units and select the most appropriate one.
- looking for possible characteristic subgroups related to the disease status;
- characterizing each subgroup with the estimates of the cluster-specific parameters.

**4.3. Empirical findings from mixture models fitted to LFPD data.** Considering all 67 available orderings we fitted mixtures of DB with  $d = d_K$  (DBmix), mixtures of PL with both forward and backward reference order (PLmix- $\rho_1$  and PLmix- $\rho_2$ ) and mixtures of EPL (EPLmix), the novel model we presented in section 2.3 where  $\rho$  is a parameter to be inferred. All mixtures have been implemented with a number of components varying from  $G = 1$  to  $G = 7$ . Of course, the case  $G = 1$  coincides with the assumption that observations come from a homogeneous

population without an underlying group structure. We separately applied the mixture models to the ranking of absorbance levels relative to the  $K = 11$  partially overlapping protein fragments as well as to the  $K = 11 + 2$  binding probes (spots), when we additionally included the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).

Focusing on the BIC for  $G = 1$  compared to  $G > 1$ , the MLE of the DBmix provided an overall evidence in favor of a heterogeneity when both  $K = 11$  or  $K = 13$  binding probes are considered. We highlighted a remarkable decreasing behavior for the associated BIC, which persists when fitting is carried out up to  $G = 10$  components as shown in Table 2.

[Table 2 about here.]

[Figure 5 about here.]

Indeed, fitting DBmix with an increasing number of groups pointed out a particular feature of the DB, probably due to the sparse nature of LFPD data. We remind, in fact, that in the present application the sample size is small w.r.t. to the cardinality ( $11!$  or  $13!$ ) of the discrete ranking space. As the value of  $G$  in the DBmix increases, some components start to represent just a single observation. This can be explained, perhaps, by the fact that, once the modal ranking  $\sigma$  has been fixed, DB has only one remaining parameter fitting the shape of the uncertainty. It follows that for these components the concentration parameter  $\lambda$  is typically estimated as a very high value. This behavior, of course, could make the DBmix model not sufficiently parsimonious and promising in some sparse-data situations, because its fit would lead to a more sparse clustering of the observations and to a less enlightening inferential findings. When stage-wise models have been fitted to the LFPD data, we found again evidence in favor of the heterogeneous structure, as indicated in Table 3; in the case  $K = 11$  all types of mixtures consistently identify 4 groups in the sample, whereas one additional component is selected by both PLmix- $\rho_1$  and EPLmix when also the control probes are included in the ordered sequences. Bold BIC values in Table 3 indicate the EPLmix as the best model and are, in both cases, significantly smaller than those of the competing mixtures. Indeed, this is constantly true for every considered dimension  $G$  of the mixture, as elucidated by Figure 5. This proves the successful introduction of the discrete parameter  $\rho$  which drastically improves the fitting of the data. Moreover, the EPLmix exhibits a good accuracy in the discrimination of sample units w.r.t. the real disease status.

[Table 3 about here.]

The two resulting clusterings agree with the most relevant distinction of the real disease status (healthy/non healthy), as pointed out in Tables 4(a) and 4(b). Specifically, collapsing also the model-based group membership into this basic bipartition we recognize that healthy subjects are well isolated with only 1 or 2 false positive cases, whereas for diseased patients we have 7 misclassifications when  $K = 11$  but only 2 with the addition of the control spots, see Tables 4(a) and 4(b). As expected, the inclusion of the positive and negative controls produced a fruitful discriminant evidence, suggested by the global reduction of misclassifications for clusterings based on 13 ranks. Healthy patients are always modeled with two components in all mixtures fitted to the LFPD data. This hints at possibly different subtypes of healthy profiles. In fact, we easily verified that such subdivision reflects two different absorbance patterns in cancer-free units, made evident in Figure 2 by

the green broken lines: one subgroup whose immune defenses essentially did not react at all to the exposition with the HER2 oncoprotein (lower panel) and those with some manifest and characterized binding profile (upper panel).

[Table 4 about here.]

On the other hand, among the selected components which can be categorized as corresponding to diseased patients, the sub-classification between EBC and MBC is only partially recovered, especially for the latter group of patients. This is proved by the presence of at least one model-based group entirely composed of EBC subjects in all fitted mixtures, whereas MBC always belong to mixed-type components.

The varying correspondence between the real cancer status and the inferred clustering structure confirms the presumed dependence of the classification results on the adopted reference ranking process  $\rho$ . Furthermore, the good agreement obtained with the EPLmix, and pointed out by Tables 4(a) and 4(b), suggests that researchers should not focus exclusively on differential epitope identification but should extend their analysis considering also more general global understanding of differential bindings. Hence, in order to characterize disease groups w.r.t. ranking profiles it is interesting to interpret the component-specific modal orderings (Table 5), derived by ordering the corresponding support parameter estimates (Figure 6). Weights and reference order estimates for the identified clusters are shown in Table 6.

[Table 5 about here.]

[Table 6 about here.]

[Figure 6 about here.]

Focusing on the analysis based on 13 binding probes, we stress that in all the best fitted models the positive control probe, labeled as Hum 12, recurrently occupies top positions in modal orderings of EBC and EBC+MBC mixture components. We remind that Hum 12 denotes the absorbance level corresponding to the entire HER2 oncoprotein. Thus, in theory, its level should reflect the total binding and it is reasonably expected to be higher than absorbance level detected in limited portions of the oncoprotein. On the other hand, immunological response in healthy patients may either be unaffected by the exposition with the HER2 oncoprotein or yield a mild binding. This means an exchangeability of binding probes in the ordering of absorbance levels, which is typical under the UM. These aspects reinforce the presence of Hum 12 in top positions as a signal that the immunological response actually occurred and hence it can be interpreted as a distinguishing feature of the unhealthy patients. It turns out that with our wildly fluctuating LFPD data it is not possible to identify a simple threshold for the raw (or normalized) binding outcome to discriminate unhealthy patients. This is better achieved using binding profiles based on rankings. Moreover, the combination of Hum 12 with the pattern (Hum 1, Hum 11, Hum 7) in top positions seems to characterize mixed (EBC+MBC) diseased groups, such as the first and the fourth components in PLmix- $\rho_1$ , the third one in PLmix- $\rho_2$  and the first one in EPLmix. In fact the protein fragments Hum 1, Hum 11 and 7 were already recognized in (Gabrielli et al., 2013) as the relevant epitopes. In EBC-specific components, similar results are valid for the fragment pair (Hum 9, Hum 3) which, together with the positive control, occupies the very first top positions, see for example the third group in PLmix- $\rho_1$  and the second one in PLmix- $\rho_2$ ; this means that for some EBC patients the binding reaction mainly



occurs in a different section of the oncoprotein, improving the discrimination of this subgroup among diseased patients. Relevant findings can be also highlighted for healthy patients. The absent or negligible immunological response observed for some of them is well described by estimated models with a component which is very close to the UM, as shown by the corresponding inferred value  $\hat{p}_g$ . In this case the modal orderings are poorly representative, so we marked them with the symbol \* in Table 5. These UM components involve prevalently HD patients. They are also included in another more characterized mixture component. The interpretation of the non-uniform component parameters suggests that some HD subjects share the epitopes Hum 1 and Hum 7 with other patients but they also have a distinctive Hum 2 in top positions; Hum 11, instead, appears in middle positions. We can also look at low absorbance patterns, if bottom ranks can be regarded as meaningful *signatures* for the problem at hand. Note, for example, that whereas Hum 10 appears consistently in last positions for almost all of the fitted components, Hum 9 seems to be a sort of “anti”-epitope signature for HD units; the same role is played by the Hum pattern (Hum 5, Hum 6) for EBC subjects. Another interesting feature regards Hum 13; it corresponds to the empty phage vector hence, theoretically, one would expect it to be associated with bottom ranks whereas this is true specifically for those groups composed for the most part of MBC units, see for example the fourth component in the PLmix- $\rho_1$ , the third one in the PLmix- $\rho_2$  and the first one in the EPLmix. Hence, a minimum absorbance level in Hum 13 could be an important indication to discriminate MBC patients, the subgroup which is only weakly characterized by the present analysis. Similar observations are valid for the case  $K = 11$  omitting, naturally, Hum 12 and Hum 13. Finally, we remark that the model selected as the best in terms of the BIC is the EPLmix, which involves 69 parameters in the case of  $K = 13$ .

**4.4. Alternative quantitative data analysis.** Now we show that our analysis based on ranked data and EPL mixture model compares favorably with a more conventional approach relying on quantitative data. We implemented the flexible mixture of multivariate normal distributions (MNorm-mix) with the R package `mclust` described in (Fraley and Raftery, 2003).

As urged in section 4.2, we must preliminarily decide whether there exists a more appropriate way of transforming and rescaling the original quantitative measures. Because a consolidated normalization method is lacking for this type of experiments, we worked with 3 alternative reasonable options: original raw data, the log-transformed absorbances and the rescaled log-transformed absorbances so that the individual average log-absorbance of all the considered spots is null for each patient. Results derived from the quantitative analysis are very different according to which measurement scale is used in the input data. In fact, only with the raw data the best fitting mixture model provides evidence in favor of an heterogeneous model, namely a mixture with  $G = 3$  components. However, as shown in Table 7(a), the correspondence with the known disease status is poorer than the one obtained with the ranking-based analysis. In all other cases the MNorm-mix model selected the single component homogeneous model as best fitting. However, if one forces the model to be fitted as heterogeneous then a variable number of groups is selected, ranging from 4 to 7. Indeed, the best classification that one can obtain with a MNorm-mix fitted to the rescaled log-transformed absorbances of all the 13 Hum has a very good agreement with the three disease subgroups, as shown in

Table 7(b). However, we stress that this model is not selected as the best fitting in terms of BIC and yields a more scattered clustering. Moreover, this model requires 117 parameters and hence it is less parsimonious and would be more difficult to interpret than the best fitting mixture for ranked data.

[Table 7 about here.]

## 5. CONCLUDING REMARKS AND FUTURE DEVELOPMENTS

In the present work we presented a novel extension of the popular and widely-used Plackett-Luce model relaxing the standard assumption of forward ranking elicitation and detailed its estimation in the MLE framework. We verified the usefulness of the EPL with a successful application to the real LFPD data set from a bioassay experiment, comparing its performance w.r.t. alternative and more standard probability distributions for rankings. Specifically, taking into account the heterogeneous origin of the sample units, we considered several parametric models in a mixture model setting. Inferential results of our mixture modeling approach pointed out a good capability of the absorbance rankings to fit heterogeneous and wilding fluctuating binding data and a good accuracy in discriminating the actual disease status. Interestingly, an almost UM component has been estimated from the data. Differently from previous applications in the literature, where the UM component was introduced to fit outliers/untypical observations, for the LFPD data such a component does not have the marginal role to model noise in the sample but has a precise interpretation to characterize healthy patients. The utility of the ranking-based analysis for epitope mapping experiments is reinforced by the possibility to partially overcome difficulties related to the choice of the preliminary normalization, needed for the raw quantitative absorbance profiles. Additionally, the fitted model turns out to be more parsimonious than alternative quantitative analyses for the present multivariate setting and exhibits an interesting interpretation, unaffected by ad-hoc monotone pre-processing transformations of the original raw data. Hence, our work suggests that even when quantitative data are available in a bioassay experiment, statistical analysis of the underlying ordinal information may provide a useful and more robust tool for the description of the outcomes. Cluster-specific parameter estimates, characterizing groups of patients, are very useful to construct an epitope mapping profile, i.e., to identify protein fragments whose binding can be related to the disease development and to detect spots relevant for possible classification/prediction purposes. Moreover, the significantly improved fit for the present application obtained with the more general EPL class can be explained with the fact that our proposal accounts for the absence of a natural and *a priori* known reference order of the binding mechanism and consequently allows to capture the discriminant contribution of all positions. This suggests that the understanding of the binding outcomes should not be limited to the use of the standard forward PL.

A first natural way to develop further our work consists in implementing the EPL mixture model in a Bayesian framework, in order to allow the incorporation of pre-experimental information in the analysis. This extension could benefit from the conjugacy of the PL with the Gamma prior distribution, already exploited for the Bayesian inference in (Guiver and Snelson, 2009) and (Caron and Doucet, 2012) but restricted to the homogenous population case. Another interesting direction

could be that of setting up a flexible framework to integrate the use of mixed-type (ordinal and quantitative) data with the possible inclusion of individual covariates.

#### ACKNOWLEDGEMENTS

We are deeply grateful to Augusto Amici and his research group for providing LFPD data and many insightful discussions on the biological foundations of the experimental outcomes. We also thank prof. Thomas Brendan Murphy and Isobel Claire Gormley for providing their C code and helpful hints.

#### APPENDIX

We prove here the presence of distributions on orderings in our novel EPL family which are not members of the canonical PL. For this purpose, let us remind that the PL implies the *independence of irrelevant alternatives* (IIA), stating that the probability ratio to select an item over another is unaffected by the preferences towards the other alternatives in the choice set, see (Luce, 1959). Equivalently, one can say that in a PL the choice probability ratio between two items is constant over all stages as long as such alternatives are both still available. In the  $K = 3$  case the IIA lemma translates into the following set of conditions on the probabilities  $q_{\pi^{-1}} = \mathbf{P}(\pi^{-1})$  of each possible ordering

$$(5.1) \quad \begin{aligned} \frac{q(1,2,3)}{q(1,3,2)} &= \frac{q(2,1,3) + q(2,3,1)}{q(3,1,2) + q(3,2,1)} \\ \frac{q(2,1,3)}{q(2,3,1)} &= \frac{q(1,2,3) + q(1,3,2)}{q(3,1,2) + q(3,2,1)} \\ \frac{q(3,1,2)}{q(3,2,1)} &= \frac{q(1,2,3) + q(1,3,2)}{q(2,1,3) + q(2,3,1)} \end{aligned}$$

and they have to be simultaneously satisfied for a generic ranking distribution to belong to the forward PL. Now, let us consider the EPL with fixed  $\rho = (2, 1, 3)$  and the generic induced probability function on random orderings given by

$$(5.2) \quad \begin{pmatrix} q(1,2,3) & q(1,3,2) & q(2,1,3) & q(2,3,1) & q(3,1,2) & q(3,2,1) \\ \frac{p_2 p_1}{1-p_2} & \frac{p_3 p_1}{1-p_3} & \frac{p_1 p_2}{1-p_1} & \frac{p_3 p_2}{1-p_3} & \frac{p_1 p_3}{1-p_1} & \frac{p_2 p_3}{1-p_2} \end{pmatrix}.$$

Substituting (5.2) in (5.1) and solving w.r.t.  $p$  one obtains as unique solution  $\underline{p} = (1/3, 1/3, 1/3)$ , meaning that the two model classes can share only the UM. This formally shows what has been hinted at in (Fligner and Verducci, 1988) on the possibility to define new ranking models relaxing the forward hypothesis. To give an intuition about the types of ranking distributions that are not covered by the traditional PL, let us consider the EPL with parameter configuration  $\rho = (2, 1, 3)$  and  $\underline{p} = (1 - 2\epsilon, \epsilon, \epsilon)$  where  $\epsilon \rightarrow 0$ . The corresponding probability function over the six possible orderings has two equally supported modes on the sequences with item 1 ranked second capturing almost the total mass, as shown in Figure 7(a). This represents a distribution that can not be obtained with any parameter specification from the forward PL. In fact, the suitable calibration of the support parameters can lead only to degenerate marginal choices of item 1 for the first and the last rank, see Figures 7(b) and 7(c). Therefore, the introduction of the parameter  $\rho$  running in the permutation space allows to overcome this asymmetry among ranks.

[Figure 7 about here.]

## REFERENCES

- William Benter. Computer based horse race handicapping and wagering systems: A report. In Donald B. Hausch, Victor S.Y. Lo, and William T. Ziemba, editors, *Efficiency of racetrack betting markets*, pages 183–198. Academic Press, 1994.
- Ludwig M. Busse, Peter Orbanz, and Joachim M. Buhmann. Cluster analysis of heterogeneous rank data. In Zoubin Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning – ICML 2007*, pages 113–120. Omnipress, 2007.
- François Caron and Arnaud Doucet. Efficient bayesian inference for generalized bradley-terry models. *J. Comput. Graph. Statist.*, 21(1):174–196, 2012. ISSN 1061-8600. doi: 10.1080/10618600.2012.638220.
- Douglas E. Critchlow. *Metric methods for analyzing partially ranked data*, volume 34 of *Lecture Notes in Statistics*. Springer-Verlag, 1985. ISBN 3-540-96288-3. doi: 10.1007/978-1-4612-1106-8.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. ISSN 0035-9246. with discussion.
- Michael A. Fligner and Joseph Stephen Verducci. Distance based ranking models. *J. Roy. Statist. Soc. Ser. B*, 48(3):359–369, 1986. ISSN 0035-9246.
- Michael A. Fligner and Joseph Stephen Verducci. Multistage ranking models. *J. Amer. Statist. Assoc.*, 83(403):892–901, 1988. ISSN 0162-1459.
- Chris Fraley and Adrian E. Raftery. Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. *J. Classification*, 20(2): 263–286, 2003. ISSN 0176-4268. doi: 10.1007/s00357-003-0015-3.
- Federico Gabrielli, Roberto Salvi, Chiara Garulli, Cristina Kalogris, Serena Arima, Luca Tardella, Paolo Monaci, Serenella Maria Pupa, Elda Tagliabue, Maura Montani, Elena Quaglino, Claudia Curcio, Cristina Marchini, and Augusto Amici. Identification of relevant conformational epitopes on the her2 oncoprotein by using large fragment phage display (lfpd). *PlosONE*, 8(3), 2013. doi: 10.1371/journal.pone.0058358.
- Isobel Claire Gormley and Thomas Brendan Murphy. Analysis of irish third-level college applications data. *J. Roy. Statist. Soc. Ser. A*, 169(2):361–379, 2006. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2006.00412.x.
- Isobel Claire Gormley and Thomas Brendan Murphy. Exploring voting blocs within the irish electorate: a mixture modeling approach. *J. Amer. Statist. Assoc.*, 103(483):1014–1027, 2008a. ISSN 0162-1459. doi: 10.1198/016214507000001049.
- Isobel Claire Gormley and Thomas Brendan Murphy. A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Stat.*, 2(4):1452–1477, 2008b. ISSN 1932-6157. doi: 10.1214/08-AOAS178.
- Isobel Claire Gormley and Thomas Brendan Murphy. A grade of membership model for rank data. *Bayesian Anal.*, 4(2):265–295, 2009. ISSN 1936-0975.
- John Guiver and Edward Snelson. Bayesian inference for plackett-luce ranking models. In Léon Bottou and Michael Littman, editors, *Proceedings of the 26th International Conference on Machine Learning – ICML 2009*, pages 377–384. Omnipress, 2009. ISBN 978-1-60558-516-1.
- David R. Hunter. Mm algorithms for generalized bradley-terry models. *Ann. Statist.*, 32(1):384–406, 2004. ISSN 0090-5364. doi: 10.1214/aos/1079120141.

- David R. Hunter and Kenneth Lange. A tutorial on mm algorithms. *Amer. Statist.*, 58(1):30–37, 2004. ISSN 0003-1305. doi: 10.1198/0003130042836.
- Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.*, 9(1):1–59, 2000. ISSN 1061-8600. doi: 10.2307/1390605. With discussion, and a rejoinder by Hunter and Lange.
- R. D. Luce. *Individual choice behavior: A theoretical analysis*. John Wiley & Sons Inc., 1959.
- John I. Marden. *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman Hall, 1995. ISBN 0-412-99521-2.
- Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000. ISBN 0-471-00626-2. doi: 10.1002/0471721182.
- P. D. McNicholas, T. B. Murphy, A. F. McDaid, and D. Frost. Serial and parallel implementations of model-based clustering *via* parsimonious gaussian mixture models. *Comput. Statist. Data Anal.*, 54(3):711–723, 2010. ISSN 0167-9473. doi: 10.1016/j.csda.2009.02.011.
- Thomas Brendan Murphy and Donal Martin. Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.*, 41(3-4):645–655, 2003. ISSN 0167-9473. doi: 10.1016/S0167-9473(02)00165-2.
- Robin L. Plackett. The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 24(2):193–202, 1975. ISSN 0035-9254.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL <http://www.R-project.org/>. ISBN 3-900051-07-0.
- Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978. ISSN 0090-5364.
- Arthur Richard Silverberg. *Statistical models for  $q$ -permutations*. ProQuest LLC, Ann Arbor, MI, 1980. Thesis (Ph.D.)—Princeton University.

## LIST OF FIGURES

- 1 1-D scheme of the HER2 oncoprotein and its segmentation into 11 partially overlapping fragments (Hum) employed in the LFPD bioassay experiment. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control). 23
- 2 Raw absorbance profiles for the three group of patients in the LFPD study: HD = healthy (green), EBC = diagnosed with breast cancer in an early stage (red), MBC = diagnosed with metastatic breast cancer (blue). Each broken line represents the absorbance levels in the HER2 oncoprotein fragments (Hum) of a single experimental unit. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control). 24
- 3 Mean-centered log-absorbance profiles for the three group of patients in the LFPD study: HD = healthy (green), EBC = diagnosed with breast cancer in an early stage (red), MBC = diagnosed with metastatic breast cancer (blue). Each broken line represents the mean-centered log-absorbance levels in the HER2 oncoprotein fragments (Hum) of a single experimental unit. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control). 25
- 4 Image plots of the first-order marginal matrices for the three groups of patients in the LFPD study: HD = healthy (left), EBC = diagnosed with early stage breast cancer (center), MBC = diagnosed with metastatic breast cancer (right). Upper panel refers to the data with  $K = 11$  protein fragments whereas the lower one concerns the  $K = 13$  case with the addition of Hum 12 and Hum 13, indicating respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control). The Borda ordering  $\overline{\pi}^{-1}$  lists items taken in order from the highest to the lowest mean rank. 26
- 5 BIC trends resulting from the MLE of the PLmix- $\rho_1$ , the PLmix- $\rho_2$  and the EPLmix on the LFPD data with a varying number  $G$  of mixture components, when either  $K = 11$  (left) or  $K = 13$  (right) binding probes are included in the ranking. The symbol  $\star$  indicates the minimum BIC values for the final selection of the number of groups. 27
- 6 Support parameter estimates represented via mosaic plots for the best PLmix- $\rho_1$ , PLmix- $\rho_2$  and EPLmix fitted to the LFPD data. Bar widths are proportional to the group weights. Upper panel refers to the data with  $K = 11$  protein fragments whereas the lower one concerns the  $K = 13$  case with the addition of Hum 12 and Hum 13, indicating respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control). 28
- 7 Examples of EPL (left) and PL (center and right) distribution functions on random orderings. 29

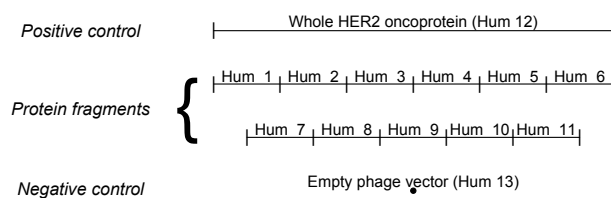


FIGURE 1. 1-D scheme of the HER2 oncoprotein and its segmentation into 11 partially overlapping fragments (Hum) employed in the LFPD bioassay experiment. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).

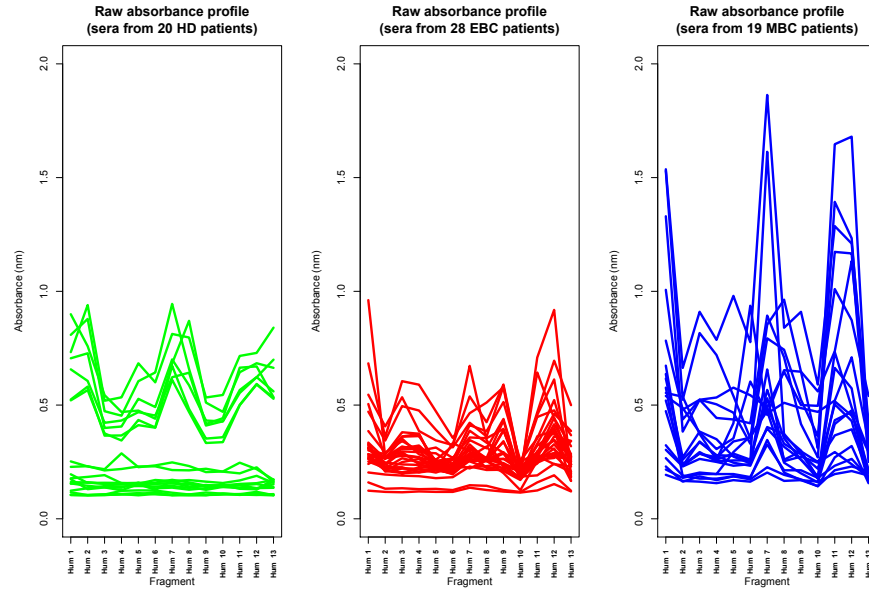


FIGURE 2. Raw absorbance profiles for the three group of patients in the LFPD study: HD = healthy (green), EBC = diagnosed with breast cancer in an early stage (red), MBC = diagnosed with metastatic breast cancer (blue). Each broken line represents the absorbance levels in the HER2 oncoprotein fragments (Hum) of a single experimental unit. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).



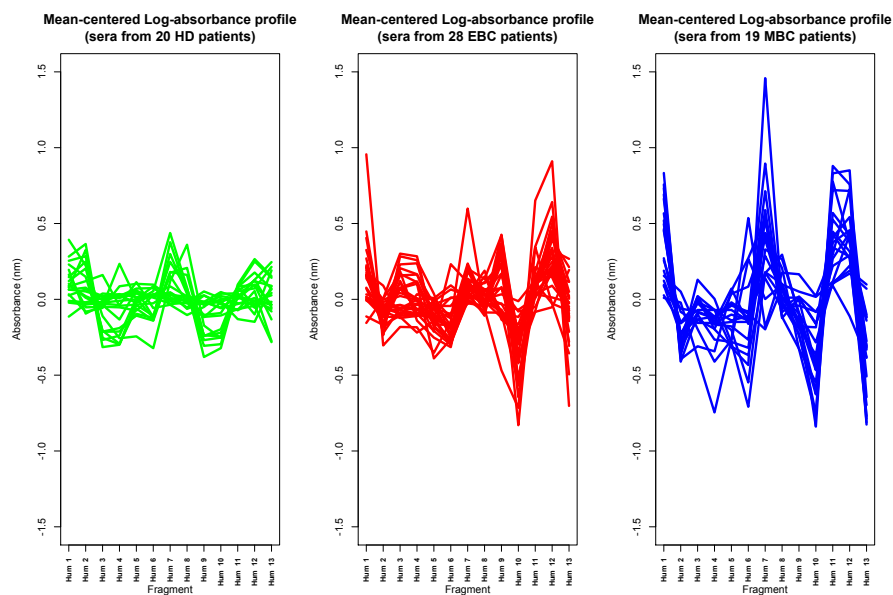


FIGURE 3. Mean-centered log-absorbance profiles for the three group of patients in the LFPD study: HD = healthy (green), EBC = diagnosed with breast cancer in an early stage (red), MBC = diagnosed with metastatic breast cancer (blue). Each broken line represents the mean-centered log-absorbance levels in the HER2 oncoprotein fragments (Hum) of a single experimental unit. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).

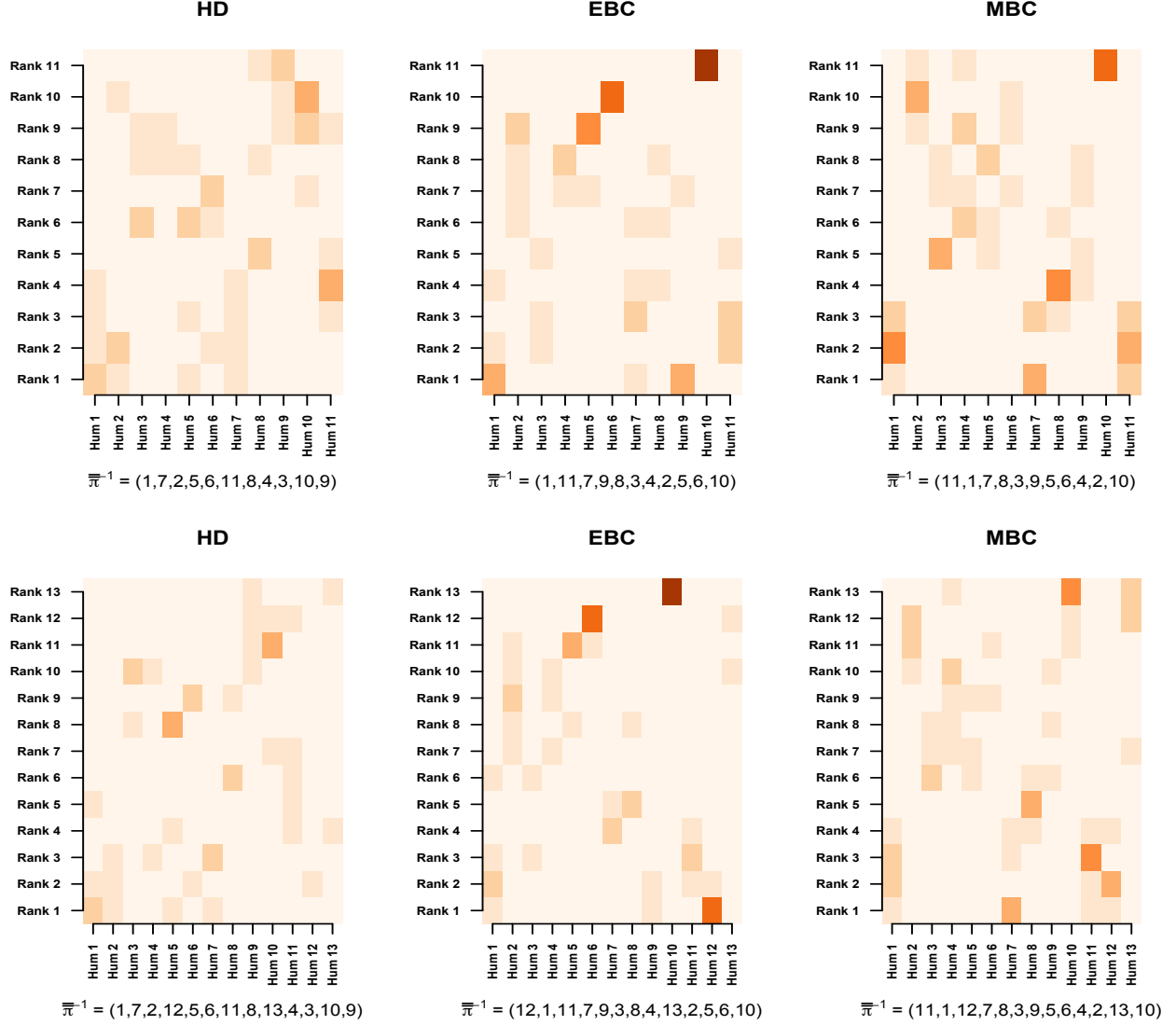


FIGURE 4. Image plots of the first-order marginal matrices for the three groups of patients in the LFPD study: HD = healthy (left), EBC = diagnosed with early stage breast cancer (center), MBC = diagnosed with metastatic breast cancer (right). Upper panel refers to the data with  $K = 11$  protein fragments whereas the lower one concerns the  $K = 13$  case with the addition of Hum 12 and Hum 13, indicating respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control). The Borda ordering  $\bar{\pi}^{-1}$  lists items taken in order from the highest to the lowest mean rank.

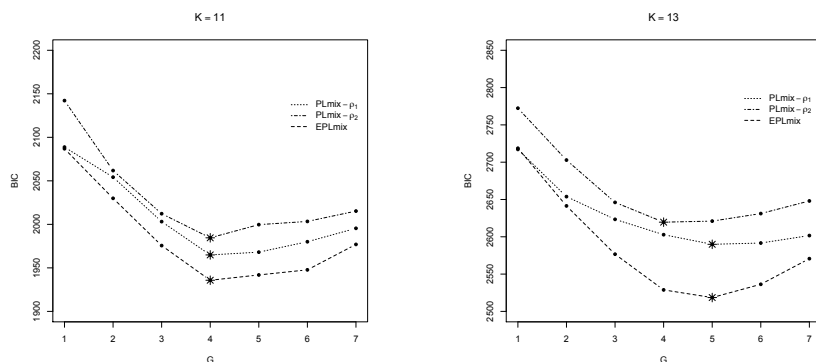


FIGURE 5. BIC trends resulting from the MLE of the PLmix- $\rho_1$ , the PLmix- $\rho_2$  and the EPLmix on the LFPD data with a varying number  $G$  of mixture components, when either  $K = 11$  (left) or  $K = 13$  (right) binding probes are included in the ranking. The symbol  $\star$  indicates the minimum BIC values for the final selection of the number of groups.

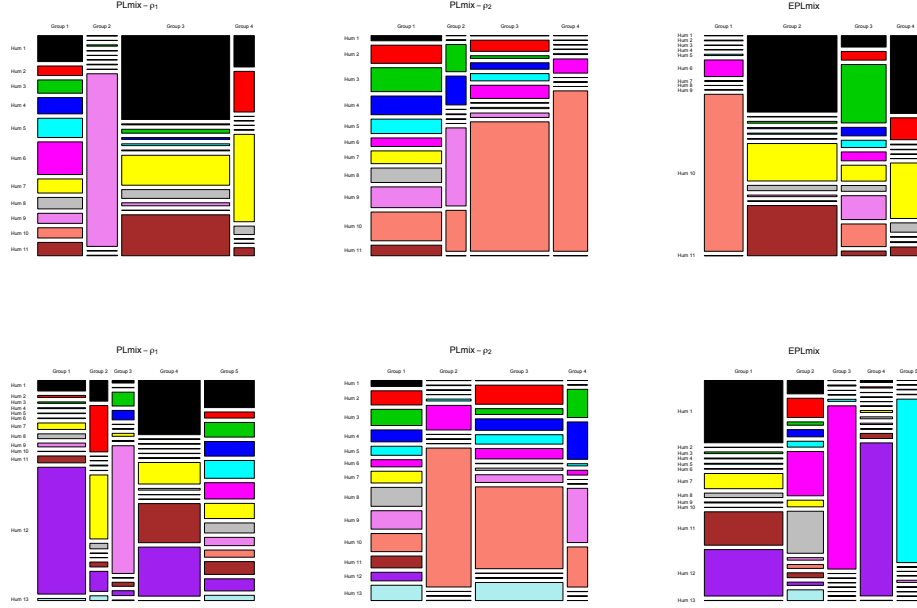


FIGURE 6. Support parameter estimates represented via mosaic plots for the best PLmix- $\rho_1$ , PLmix- $\rho_2$  and EPLmix fitted to the LFPD data. Bar widths are proportional to the group weights. Upper panel refers to the data with  $K = 11$  protein fragments whereas the lower one concerns the  $K = 13$  case with the addition of Hum 12 and Hum 13, indicating respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).

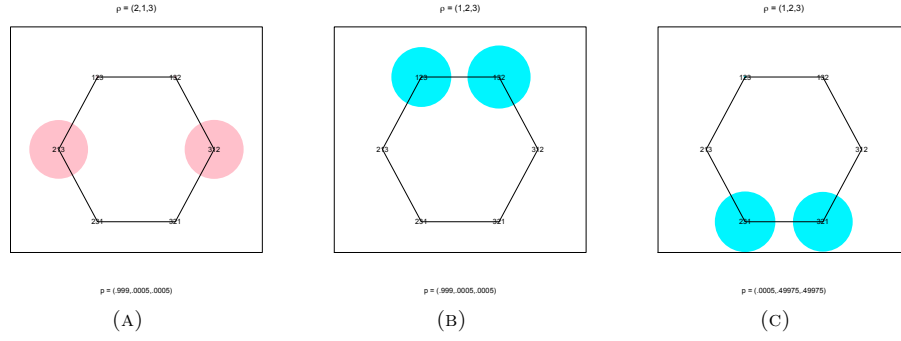


FIGURE 7. Examples of EPL (left) and PL (center and right) distribution functions on random orderings.

TABLE 1. BIC values resulting from the MLE of the PL ( $\rho_1 =$  forward and  $\rho_2 =$  backward reference order) and of the EPL on the three disease groups for a different number  $K$  of binding probes included in the ranking. Groups of patients are defined as follows: HD = healthy, EBC = diagnosed with early stage breast cancer, and MBC = diagnosed with metastatic breast cancer.

Model	$K = 11$			$K = 13$		
	HD	EBC	MBC	HD	EBC	MBC
PL- $\rho_1$	694.04	776.13	499.46	899.63	<b>1025.71</b>	658.44
PL- $\rho_2$	685.85	804.61	498.67	894.44	1039.45	652.15
EPL	<b>676.93</b>	<b>773.17</b>	<b>473.90</b>	<b>873.71</b>	1026.61	<b>630.05</b>

TABLE 2. BIC values resulting from the MLE of the DBmix on the LFPD data with a varying number  $G$  of components, when either  $K = 11$  or  $K = 13$  binding probes are included in the ranking.

	$G$									
	1	2	3	4	5	6	7	8	9	10
$K = 11$	2078.77	2003.65	1940.86	1899.33	1882.32	1863.17	1846.98	1829.81	1817.06	1798.12
$K = 13$	2700.02	2617.66	2551.38	2512.19	2483.25	2451.38	2421.71	2392.10	2366.78	2342.60

TABLE 3. BIC values and number  $G$  of components of the best PLmix- $\rho_1$ , PLmix- $\rho_2$  and EPLmix fitted to LFPD data for different number  $K$  of binding probes included in the ranking.

Mixture	$K = 11$		$K = 13$	
Model	BIC	$G$	BIC	$G$
PLmix- $\rho_1$	1964.87	4	2589.93	5
PLmix- $\rho_2$	1984.53	4	2619.60	4
EPLmix	<b>1935.77</b>	4	<b>2518.56</b>	5



TABLE 4. Correspondence between the model-based clustering derived by the MLE of the EPLmix and the true disease status of the LFPD experimental units: HD = healthy, EBC = diagnosed with early stage breast cancer and MBC = diagnosed with metastatic breast cancer.

(A) $K = 11$					(B) $K = 13$					
Disease Status	Group				Disease Status	Group				
	1	2	3	4		1	2	3	4	5
HD	0	2	10	8	HD	1	10	0	1	8
EBC	13	12	2	1	EBC	12	0	9	7	0
MBC	0	15	3	1	MBC	14	2	0	3	0

TABLE 5. Modal orderings and composition w.r.t. the real cancer status of the components identified with the best PLmix- $\rho_1$ , PLmix- $\rho_2$  and EPLmix fitted to LFPD data, for a different number  $K$  of binding probes included in the ranking. “D.C.” stands for “disease composition” listing sequentially the number of HD = healthy, EBC = diagnosed with early stage breast cancer and MBC = diagnosed with metastatic breast cancer patients in each group.

Mixture model	$K = 11$			$K = 13$		
	$g$	D.C.	$\hat{\sigma}_g^{-1}$	$g$	D.C.	$\hat{\sigma}_g^{-1}$
PLmix- $\rho_1$	1	(11, 1, 3)	(6, 1, 5, 4, 7, 11, 3, 8, 10, 9, 2)*	1	(2, 11, 3)	(12, 1, 11, 7, 8, 9, 2, 13, 3, 4, 6, 5, 10)
	2	(0, 10, 0)	(9, 3, 7, 11, 4, 1, 8, 2, 5, 6, 10)	2	(6, 0, 0)	(7, 2, 1, 12, 8, 11, 13, 5, 6, 10, 4, 3, 9)
	3	(3, 17, 15)	(1, 11, 7, 8, 3, 9, 4, 5, 2, 6, 10)	3	(0, 7, 0)	(9, 3, 4, 12, 11, 7, 1, 8, 13, 2, 5, 6, 10)
	4	(6, 0, 1)	(7, 2, 1, 8, 11, 5, 6, 4, 10, 3, 9)	4	(0, 7, 14)	(1, 12, 11, 7, 8, 3, 5, 9, 4, 6, 2, 13, 10)
				5	(12, 3, 2)	(1, 5, 6, 7, 3, 4, 11, 12, 8, 9, 10, 2, 13)*
PLmix- $\rho_2$	1	(14, 5, 3)	(1, 6, 11, 7, 5, 8, 2, 4, 9, 3, 10)*	1	(12, 2, 2)	(1, 6, 12, 5, 7, 11, 4, 2, 13, 3, 10, 9, 8)*
	2	(6, 0, 1)	(7, 2, 1, 8, 11, 5, 6, 3, 4, 10, 9)	2	(0, 15, 0)	(12, 9, 3, 7, 4, 11, 1, 13, 8, 2, 5, 6, 10)
	3	(0, 12, 15)	(1, 7, 11, 8, 3, 9, 4, 5, 2, 6, 10)	3	(1, 11, 17)	(12, 1, 11, 7, 8, 3, 9, 5, 6, 4, 13, 2, 10)
	4	(0, 11, 0)	(9, 3, 4, 7, 11, 1, 8, 2, 5, 6, 10)	4	(7, 0, 0)	(7, 2, 1, 12, 8, 13, 11, 5, 6, 3, 4, 10, 9)
EPLmix	1	(0, 13, 0)	(9, 8, 1, 3, 11, 7, 2, 4, 5, 6, 10)	1	(1, 12, 14)	(12, 1, 11, 7, 8, 3, 4, 9, 5, 6, 2, 13, 10)
	2	(2, 12, 15)	(1, 11, 7, 8, 9, 3, 4, 5, 6, 2, 10)	2	(10, 0, 2)	(5, 2, 11, 4, 3, 6, 10, 7, 8, 9, 12, 1, 13)
	3	(10, 2, 3)	(5, 4, 11, 1, 6, 3, 10, 2, 9, 7, 8)	3	(0, 9, 0)	(9, 12, 11, 3, 1, 4, 7, 2, 13, 8, 5, 6, 10)
	4	(8, 1, 1)	(7, 2, 1, 8, 11, 5, 6, 4, 10, 9, 3)	4	(1, 7, 3)	(12, 9, 1, 11, 13, 3, 8, 7, 2, 4, 5, 6, 10)
				5	(8, 0, 0)	(11, 2, 1, 6, 12, 8, 13, 5, 7, 10, 4, 3, 9)

Note: The symbol \* indicates mixture components which are very close to the UM.

TABLE 6. Mixture weights and reference order estimates of the best PLmix- $\rho_1$ , PLmix- $\rho_2$  and EPLmix fitted to LFPD data, for a different number  $K$  of binding probes included in the ranking.

Mixture model	$K = 11$		$K = 13$	
	$g$	$\hat{\omega}_g$	$g$	$\hat{\omega}_g$
PLmix- $\rho_1$	1	.22	$\rho_1$	
	2	.15	$\rho_1$	
	3	.53	$\rho_1$	
	4	.10	$\rho_1$	
			$\rho_1$	
PLmix- $\rho_2$	1	.35	$\rho_2$	
	2	.10	$\rho_2$	
	3	.39	$\rho_2$	
	4	.16	$\rho_2$	
			$\rho_2$	
EPLmix	1	.19 (11, 10, 9, 7, 8, 4, 2, 3, 6, 5, 1)	1	.39 (2, 1, 3, 4, 5, 6, 8, 9, 7, 10, 11, 12, 13)
	2	.44 (1, 2, 3, 4, 6, 5, 7, 8, 9, 10, 11)	2	.18 (6, 9, 2, 12, 13, 4, 8, 1, 3, 7, 11, 5, 10)
	3	.22 (6, 9, 7, 10, 4, 5, 8, 2, 1, 11, 3)	3	.14 (12, 11, 8, 10, 9, 5, 7, 6, 3, 4, 2, 1, 13)
	4	.15 (3, 1, 2, 4, 5, 9, 11, 10, 8, 7, 6)	4	.17 (1, 4, 3, 7, 8, 2, 9, 5, 6, 10, 12, 13, 11)
			5	.12 (8, 13, 12, 10, 11, 1, 6, 7, 4, 5, 9, 2, 3)

TABLE 7. Correspondence between the model-based clustering derived by the MLE of the MNorm-mix and the true disease status of the LFPD experimental units: HD = healthy, EBC = diagnosed with early stage breast cancer and MBC = diagnosed with metastatic breast cancer.

(A)		Raw		
LFPD		data		
with 11 Hum				
		Group		
Disease Status		1	2	3
HD		7	9	4
EBC		3	2	23
MBC		9	0	10

(B)		Rescaled log-transformed						
LFPD data		with 13 Hum						
		Group						
Disease Status		1	2	3	4	5	6	7
HD		7	11	2	0	0	0	0
EBC		0	0	13	1	10	4	0
MBC		0	0	3	5	0	9	2